Car Price Data Cleaning for Data Analysis and Machine learning Report

Python Project Data Cleaning: Car Price Data Cleaning for Data Analysis and Machine learning

Data Cleaning and Machine Learning Report

1. Project Objective

The primary objective of this project is to clean and preprocess car price data to ensure accuracy and usability for further data analysis and machine learning applications. This includes handling missing values, feature engineering, and standardizing numerical fields.

2. Key Performance Indicators (KPIs) & Questions

- **KPIs:**
  - Percentage of missing values before and after cleaning
  - Total number of records before and after preprocessing
  - Standardization success rate for the car price column
- **Questions Addressed:**
  - What are the major data quality issues in the dataset?
  - How can car price values be converted into a uniform format?
  - What transformations are necessary for effective analysis and modeling?

3. Process of Data Cleaning

Step 1: Data Loading & Initial Inspection

- Loaded dataset from car_price.csv
- Checked dataset shape and previewed initial records
- Identified data types of all columns

Step 2: Handling Missing Values

- Counted missing values in each column
- Calculated missing value percentages
- Removed rows containing missing values

- Rechecked dataset for any remaining null values

Step 3: Feature Engineering

- Extracted Car Company Name:
  - Derived the car manufacturer from the car_name column by splitting text
  - Created a new column car_company_name
- Standardized Car Price Column:
  - Removed commas and formatted values
  - Converted car prices expressed in 'Lakh' and 'Crore' to numerical values
  - Ensured data type was set to float64

4. Data Analysis & Machine Learning Preparation

- Conducted exploratory data analysis (EDA) to identify trends and relationships
- Prepared cleaned data for machine learning by encoding categorical variables
- Normalized and scaled numerical data for better model performance
- Split dataset into training and testing sets for machine learning models

5. Project Insights (Key Findings)

- The dataset initially contained missing values, which were successfully removed.
- The car_name column was split into car_company_name and car_name to enhance clarity.
- The car_prices_in_rupee column was converted into a standardized numerical format, enabling precise comparisons.
- The cleaned data is now suitable for advanced analytics and machine learning applications.

6. Final Conclusion & Recommendations

- The dataset is now clean and ready for further data analysis and predictive modeling.
- Future improvements could include handling outliers in pricing, normalizing additional features, and enriching the dataset with external sources.
- Applying machine learning models such as regression and classification can help derive deeper insights from the dataset.

This cleaning process ensures the dataset is structured optimally for deeper insights and predictive modeling.